Out-of-sample predictions from plant–insect food webs: robustness to missing and erroneous trophic interaction records

Pearse, Ian S ; Altermatt, Florian

Abstract: With increasing biotic introductions, there is a great need for predictive tools to anticipate which new trophic interactions will develop and which will not. Phylogenetic constraint of interactions in both native and novel food webs can make some novel interactions predictable. However, many food webs are sparsely sampled, or may include inaccurate interactions. In such cases, it is unclear whether modeling methods are still useful to anticipate novel interactions. We ran bootstrap simulations of host-use models on a Lepidoptera-plant data set to remove native trophic records or add erroneous records in order to observe the effect of missing or erroneous data on the prediction of interactions with novel plants. We found that the model was robust to a large amount of missing interaction records, but lost predictive power with the addition of relatively few erroneous interaction records. The loss of predictive power with missing records was due to inaccuracy in estimating phylogenetic distance between native and novel hosts. Removal of interaction records proportionally to their encounter frequency in the field had little effect on the loss of predictive power. Host-use models may have immediate value for predicting novel interactions from large, but sparsely sampled databases of trophic interactions.

DOI: https://doi.org/10.1890/14-1463.1

Ecological Applications, 25(7), 2015, pp. 1953–1961

Out-of-sample predictions from plant–insect food webs: robustness to missing and erroneous trophic interaction records

Ian S. Pearse[1,4] and Florian Altermatt[2,3]

[1]*Illinois Natural History Survey, 1816 South Oak Street, Champaign, Illinois 61820 USA*
[2]*Department of Aquatic Ecology, Eawag: Swiss Federal Institute of Aquatic Science and Technology, Überlandstrasse 133, CH-8600 Dübendorf, Switzerland*
[3]*Institute of Evolutionary Biology and Environmental Studies, University of Zurich Winterthurerstrasse 190, CH-8057 Zürich, Switzerland*

Abstract. With increasing biotic introductions, there is a great need for predictive tools to anticipate which new trophic interactions will develop and which will not. Phylogenetic constraint of interactions in both native and novel food webs can make some novel interactions predictable. However, many food webs are sparsely sampled, or may include inaccurate interactions. In such cases, it is unclear whether modeling methods are still useful to anticipate novel interactions. We ran bootstrap simulations of host-use models on a Lepidoptera–plant data set to remove native trophic records or add erroneous records in order to observe the effect of missing or erroneous data on the prediction of interactions with novel plants. We found that the model was robust to a large amount of missing interaction records, but lost predictive power with the addition of relatively few erroneous interaction records. The loss of predictive power with missing records was due to inaccuracy in estimating phylogenetic distance between native and novel hosts. Removal of interaction records proportionally to their encounter frequency in the field had little effect on the loss of predictive power. Host-use models may have immediate value for predicting novel interactions from large, but sparsely sampled databases of trophic interactions.

Key words: herbivory; host-use model; introduced species; Lepidoptera; novel interactions; predictions; trophic niche model.

Introduction

Commerce and travel have weakened the longstanding biotic barriers between continents and biogeographic regions (Mack et al. 2000). The consequences of introduced organisms on native biota are difficult to predict, but can range from innocuous to very deleterious (NAS 2002). One of the most important aspects of how a novel organism affects its colonized environment is the degree to which it develops trophic interactions with existing organisms in that environment (Maron and Vilá 2001, Levine et al. 2004, Pearse et al. 2013). As such, there is considerable interest in developing predictive methods to anticipate the trophic interactions that might develop between introduced and native species before the introduction actually happens (NAS 2002, Briese 2003, Gilbert et al. 2012, Pearse and Altermatt 2013*b*, Pearse et al. 2013).

There have been several efforts to conceptually define the factors that cause some novel trophic interactions to form while others do not (Verhoeven et al. 2009, Harvey et al. 2010, Sih et al. 2010, Pearse et al. 2013). In each of these cases, the likelihood of a novel interaction can be described by the host breadth of the exploiter, the exploitability (or, conversely, defense) of the exploited organism, and the match between those organisms. Estimates of similarity between a novel organism and native organisms at the same trophic level may be useful in predicting novel interactions, because organisms that are functionally or phylogenetically similar tend to consume or be consumed by a similar set of organisms (Cattin et al. 2004, Gómez et al. 2010). A recent approach to predicting novel herbivore–plant interactions with introduced plants used only two predictors, firstly the phylogenetic distance between an introduced plant and a native host plant and secondly herbivore host breadth on native plants, and accurately predicted most Lepidoptera interactions with nonnative plants introduced into Central Europe (Pearse and Altermatt 2013*b*). This method is based on a host-use model using the native food web that then extrapolates that model to nonnative plants (an out-of-sample prediction), whose phylogenetic relationships to native flora are known (Ebert 1991–2005, Pearse and Altermatt 2013*b*). While this method was developed for novel herbivore–plant interactions, it could be applied to a wide range of novel trophic interactions, including parasite–host interactions (e.g., Ives and Godfray 2006) pollinator–plant interactions (e.g., Rezende et al. 2007), and predator–prey interactions (e.g., Naisbit et al. 2012).

Manuscript received 31 July 2014; revised 12 February 2015; accepted 19 February 2015. Corresponding Editor: C. Gratton.
[4] E-mail: ipearse@illinois.edu

TABLE 1. Descriptive statistics of the Ebert (1991–2005) Lepidoptera–plant food web.

Statistic	Value
Total number of native plants	2123
Number of Lepidoptera	898
Native interaction records	4727
Percentage fill of native food web	0.248%
Number of native non-host plants	586
Number of native host plants	1537
Total number of nonnative plants	474
Nonnative interaction records	491
Percentage fill of nonnative food web	0.115%

Notes: We list the number of Lepidoptera species and native/nonnative plant species used or not used by Lepidoptera as hosts. Percentage fill is the number of interactions that were observed divided by possible interactions (i.e., number of plants × number of Lepidoptera).

In order to be immediately useful at a large scale, such predictive methods must be able to anticipate novel trophic interactions using information on native food webs that is currently available. Initial tests of this method were conducted on a highly sampled food web, where the vast majority of native trophic interactions were known (Pearse and Altermatt 2013*b*). However, completeness is simply not the case for most food webs, which lack records of interactions due to incomplete sampling (Goldwasser and Roughgarden 1997). This is not only the case for hyperdiverse food webs in tropical areas, but also in many temperate areas. For example, researchers have attempted to define the host breadth of herbivores and pathogens of economically valuable plants based on trophic interactions present in large U.S. Department of Agriculture plant databases (Gilbert et al. 2012). Given the scope and heterogeneous sampling of trophic interactions in these databases, it is very likely that only a fraction of true hosts of herbivores and pathogens were recorded. Nevertheless, prediction of novel trophic interactions from these types of databases is a highly valuable application, given their continent-wide scope. As such, it would be very useful to know to what degree poorly sampled or inaccurate native food webs can be used to predict novel trophic interactions.

Using the Ebert Lepidoptera–plant food web (Ebert 1991–2005, Pearse and Altermatt 2013*b*) from Central Europe, we tested how missing and erroneous data affect predictions of novel trophic interactions. We deleted interaction records or added erroneous interaction records in increasing numbers in order to estimate the effects of missing and erroneous native trophic data on the prediction of novel trophic interactions. Using this technique, we asked the following related questions: First, how robust are novel host predictions to missing interaction records? Second, are novel host predictions more robust to missing or erroneous interaction records? Third, is the loss of predictive power with decreased sampling due mainly to poor parameterization of the native trophic model or due to poor extrapolation to novel hosts? Finally, do records of interactions that are rarely encountered have less of a bearing on the prediction of novel interactions than records of common interactions?

One reason why sampling of native interactions may be particularly important in predicting novel interactions is that native interactions are used twice in making those predictions. They are first used in the model parameterization step to define similarity among native hosts, and then used in the extrapolation step to define similarity between native and novel hosts. Up to now, it has been unclear which of these steps is more susceptible to missing interaction records. The approach used here allowed us not only to identify the robustness of predictions on novel trophic interactions, but also to disentangle the relative importance of the interaction information during the different steps of the model parametrization process recommended for such predictions.

METHODS

Food web

We tested how missing and erroneous data affect out-of-sample predictions from trophic models using the Ebert Lepidoptera–plant food web (Ebert 1991–2005). The food web describes the interactions between 898 larval Lepidoptera species (caterpillars of moths and butterflies) and their 1537 host plants in the German state Baden-Wuerttemberg in Central Europe (35 751 km^2). The vast majority of interaction records between Lepidoptera and host plants were compiled from a single, extensive monograph (Ebert 1991–2005), and a few additional records were added from other monographs from similar regions as well as our own personal observations (Koch and Heinicke 1991, Altermatt et al. 2006). The structure and sampling of this food web have been described elsewhere (Altermatt 2010, Altermatt and Pearse 2011, Pearse and Altermatt 2013*a*, *b*). We added information on all 586 native and nonnative plants that do not interact with any of the Lepidoptera from the complete plant list from Baden-Wuerttemberg (Bundesamt für Naturschutz 2010), making up a total of 2123 plant species considered, 474 of which have been introduced to Central Europe (Table 1). The host records analyzed here are very similar to those presented in our past work (Altermatt 2010, Altermatt and Pearse 2011, Pearse and Altermatt 2013*a*, *b*).

In our simulations, we treat the Ebert Lepidoptera–plant food web as being completely sampled and without erroneous records. While it is likely that there are some missing or erroneous interaction records within the food web, we believe that these are relatively few for several reasons. First, the host records are based on a very large number (~2.3 million) of observations of Lepidoptera–host plant interactions over the course of >50 years (Ebert 1991–2005), so the sampling intensity is high. Second, sampling was conducted with the goal of recording complete host records for each Lepidoptera

species, so additional effort was placed on recording rare hosts. Finally, the host records are based entirely on field observations in a natural setting and recorded by professional entomologists. Because of this, there are likely few erroneous observations in the data set and no records of interactions that are possible in a laboratory setting, but not in the wild.

In one set of simulations, we estimated encounter frequency from the Ebert Lepidoptera–plant food web. In the Ebert food web, values recorded an ordinal (1–5) estimate of the frequency of each Lepidoptera–plant interaction (Ebert 1991–2005). These scores were recorded as a single observation (1); a few isolated observations (2); several observations, and the plant may be locally or temporally of significance for the Lepidoptera species (3); many observations, and the plant may be locally or temporally of high significance for the Lepidoptera species (4); and very many observations, and the plant has a key role as a food source for the specific Lepidoptera species (5; Altermatt and Pearse 2011). In order to treat these ordinal scores numerically, we defined encounter frequency, where each ascending score was twice as likely to be observed than the previous score, resulting in values for scores of 1 (1), 2 (2), 3 (4), 4 (8), and 5 (16), where the encounter frequency is given in parentheses. While a twofold difference between classes is an arbitrary value, it codes interactions of higher scores as being far more likely encountered than interactions of lower scores. Our only use of encounter frequency values was to rank the likelihood of missing a given interaction. Trophic models were only used to predict the presence of an interaction, not its encounter frequency.

Plant phylogeny

Phylogenetic proximity to a native host was a key predictor in our trophic niche models. We used a recent supertree of Northern European vascular plants (Daphne) as an estimate of phylogenetic relationships between the 2597 native and nonnative plants that occur in Baden-Wuerttemberg (Durka and Michalski 2012). This regional phylogeny was based on the backbone of the APG III plant phylogeny (Bremer et al. 2009), with numerous clade-specific phylogenies grafted to appropriate nodes. This phylogeny contained 2484 (96%) of the plant species represented in the Ebert food web. The 113 species not included in the phylogeny were missing for one of two reasons. First, species boundaries were dealt with differently in 63 cases, largely pertaining to apomictic species complexes with poorly defined species concepts. In these cases, we grafted (i.e., added as a new branch) each of the 63 non-included species as sister to a species from the same species complex. In 50 cases, plants were not included in the phylogeny because they were ornamentals, and rarely encountered in naturalized settings. In these cases, we grafted the missing plant onto the Daphne phylogeny as polytomies at the genus or familial level. We then trimmed the modified Daphne phylogeny to the plants represented within Ebert food web.

Trophic niche model

We used a trophic niche model to simulate the associations of larval Lepidoptera with native host plants and to extrapolate from this model in order to predict their interactions with novel host plants (Pearse and Altermatt 2013*b*). This type of predictive technique is accomplished in three steps: native trophic model parameterization, extrapolation of the model to novel interactions, and validation of the model with information of novel interactions (Pearse et al. 2013). We provide the R functions (R Core Team 2014) used to parameterize a native trophic model and to extrapolate from that model to predict novel interactions, as used here (Pearse and Altermatt 2013*b*).

Model parameterization of native interactions.—We used a k-fold procedure to split data on native Lepidoptera–plant interactions into five ($k = 5$) random partitions, where each partition contained all information for one-fifth of the native plant records in the interaction matrix. In a previous test of different numbers of data partitioning, the number of partitions had little effect on the model parameters (Pearse and Altermatt 2013*b*). For each of the five partitions, we treated four partitions as calibration data and the fifth as evaluation data (Peterson et al. 2011:114, 274). We defined two predictors of herbivore host use, the number of native hosts of an herbivore (H) and phylogenetic distance (S), the minimum branch length separating a plant in the evaluation data partition from any host plant in the calibration data partition. We parameterized a generalized linear model (GLM) with parameters (m), where the binomial response variable of an interaction (I) between an herbivore (h) and plant (p) with number of hosts (s) was defined as

$$I_{\text{hp}} = m_{\text{h}} \times H + m_{\text{s}} \times S + m_{\text{hs}} \times S \times H.$$

Model extrapolation to novel interactions.—We averaged the parameters from the five data partitions used for the parametrization of native interactions. These averaged model parameters were then used in conjunction with values of number of native hosts of an herbivore, and phylogenetic distance between a nonnative plant and native hosts to project host use onto interactions with nonnative plants.

Validation of predictions of novel interactions.—We validated the out-of-sample prediction of novel host use (native herbivore, introduced plant) from our native trophic niche model (native herbivore, native plant) by calculating the area under the curve (AUC) of the receiver-operating characteristic (ROC) curve. This approach plots the cumulative proportion of true positive predictions against the cumulative proportion of false positive predictions (Krzanowski and Hand 2009). An uninformative model will result in an ROC

curve with an AUC of 0.5, while a perfectly predictive model will result in an ROC curve with an AUC of 1.

Bootstrapping approaches

Starting with the true set of native Lepidoptera–plant interactions ($n = 4727$) and absences of interactions between Lepidoptera and plants ($n = 1\,904\,331$) in the Ebert data set, we either removed true interactions or added erroneous interactions in the place of true absences. Records were removed either randomly or inversely proportional to their encounter frequency. In the first simulations, we removed interaction records randomly, stepwise increasing the number of records removed by 20 (0.4% of total interaction records) until there were 27 interaction records left in the native food web (i.e., 235 steps in total). Each sampling was conducted five times with replacement. In scenarios in which a lepidopteran did not have any recorded hosts, its interaction probability with all plants was assigned the global mean interaction probability. We then added false positive interaction records to the native data set in increments of 10 000 (0.5% of total absences) until the data set was saturated with false positive interactions. Due to the length of time for each simulation with high numbers of false positive interactions, we conducted each bootstrap simulation only once. In the false-positive procedure, we added tens of thousands of incorrect records to our native food web, but this rate of error is unlikely to occur in real food webs. Because of this, we reran our simulations with fewer records removed or added (between 0 and 2000 in increments of 20). In this case, each bootstrap sample was conducted five times with replacement. In order to assess the additivity of missing and erroneous data on predictions of novel interactions, we conducted 20 bootstrap simulations where 2000 records were added, removed, or both added and removed from the native food web. Because our trophic model uses the native food web in both model parameterization and model extrapolation, we determined which of these two stages was more sensitive to missing interaction records. To do this, we compared three types of interaction record removal: (1) a native food web with missing records in either model parameterization, but a full native food web for model extrapolation, (2) a full native food web for model parameterization, but one with missing records for model extrapolation, or (3) a food web with missing records for both parameterization and extrapolation. Records were removed in increments of 20. Each bootstrap sample was repeated five times for each of the three types of removals. For visualization, splines were fitted through bootstrap values using localized polynomial fitting with R function loess using a smoothing parameter (α) of 0.25. Confidence intervals around the splines were visualized as standard error in loess predictions. All simulations and analyses were conducted in R version 3.1 using package ROCR (Sing et al. 2005, R Core Team 2014).

Results

Robustness to missing and erroneous data

The trophic model with complete records of native Lepidoptera–plant interactions accurately predicted novel host use with an AUC of 0.929, which corresponds to an 83% prediction of novel host use (true positives) at a 10% false positive rate. Random removal of native Lepidoptera–plant interaction records from that data set resulted in a modest decline in predictive ability (AUC; Fig. 1A). The predictive ability of a trophic model using a data set in which two-thirds of all interaction records were removed retained an AUC of 0.843 (Fig. 1A). Once roughly two-thirds of all records from the data set were removed, predictive ability declined precipitously (Fig. 1A). In contrast, the replacement of interaction absences with false-positive interactions resulted in an immediate sharp decline in predictive ability (Fig. 1B). In this case, the predictive ability of a trophic model using a data set in which 10% of records of absence of interactions were erroneously scored as interactions dropped to an AUC of 0.639 (Fig. 1B).

We compared the effect of missing vs. erroneous trophic interaction records on the predictive ability of the trophic niche model. With realistic numbers of missing or erroneous interaction observations (i.e., 0–2000), we found that erroneous data had a more negative impact than missing data on the predictive ability of the trophic niche model if less than 1500 records (31.7% of all records) were removed or erroneously added (Fig. 2). Above this number, erroneous records had less of an impact than missing records on the predictive ability of the trophic niche model. The effect of missing and erroneous data was slightly synergistic in reducing the predictive ability of the trophic niche model when 2000 records were added, removed, or both added and removed from the native food web (Fig. 3). Erroneous data added to an under-sampled food web reduced the out-of-sample predictiveness (AUC) of the host-use model to a greater degree than erroneous data added to a completely sampled food web.

Random missing records or proportional to encounter rate

It is likely that rare or cryptic interactions will more often be overlooked than common or apparent interactions. When we removed interactions inversely proportional to their encounter rate, we found that the predictive ability of the trophic niche model was similarly robust to missing data as when that data was removed randomly (Fig. 1A).

Effect of missing data on model parameters vs. out-of-sample prediction

The native food web was used in two steps of predicting novel trophic interactions: in the estimation of model parameters, and in extrapolating to novel hosts. Removal of native interaction records from only

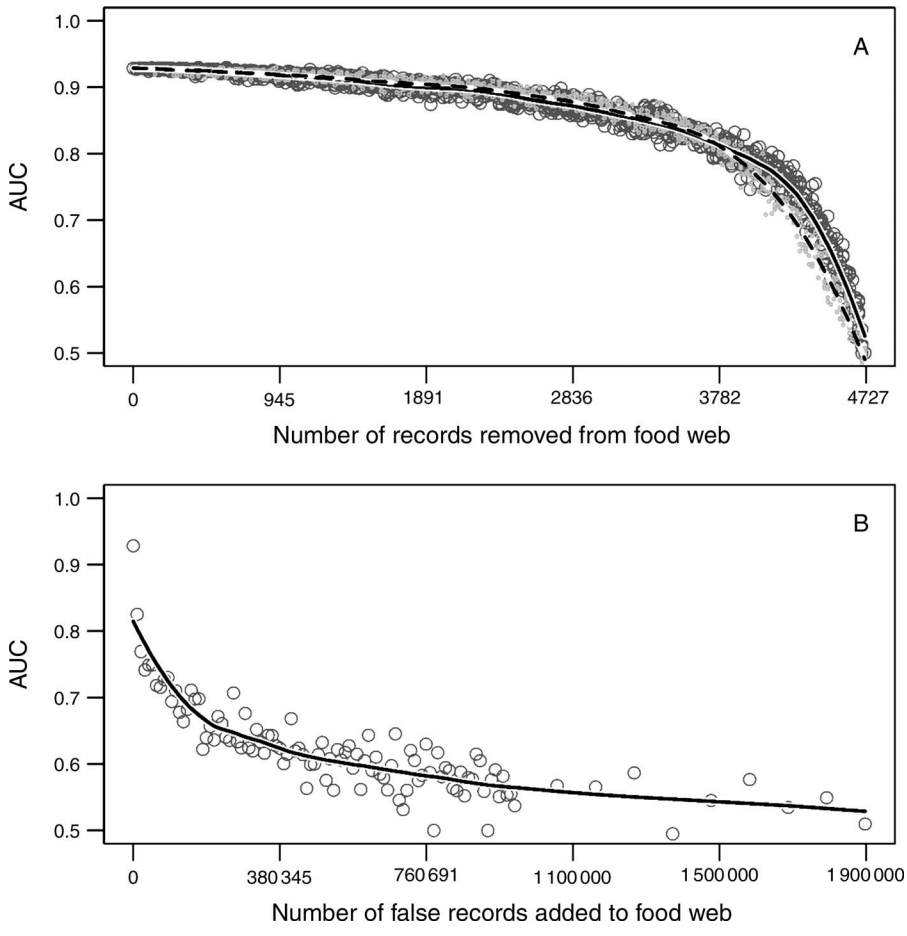

FIG. 1. Effects of missing and erroneous records on the predictiveness of a trophic model. (A) Out-of-sample (novel host) predictions (area under the curve; AUC) of trophic models onto nonnative plant–Lepidoptera interactions with increasing numbers of records removed from the native food web; numbers are shown for 0%, 20%, 40%, 60%, 80%, and 100% of records removed. Records were removed from the native food web either randomly (solid line) or inversely proportional to a categorical estimate of their encounter rate (dashed line). (B) AUC of trophic models with increasing numbers of false records of interactions added to the native food web; numbers are shown for 0%, 20%, 40%, 60%, 80%, and 100% of records removed. Open circles represent a single simulation. Lines are local-fitted polynomial splines.

the model parameterization phase of the predictive model resulted in a slower decline in predictive ability with more missing records than removal of data from both the model parameterization and extrapolation steps (Fig. 4). In contrast, removal of data from the data extrapolation step of the model resulted in a decline in predictive ability that was very similar to removal of data from both steps (Fig. 4), suggesting that model parameterization is more robust to missing interaction records than model extrapolation.

DISCUSSION

We found that a native trophic niche model of Lepidoptera–plant interactions was transferable to novel interactions between introduced plants and the same set of lepidopteran herbivores even when the native food web contained only one-third of all real Lepidoptera–plant interactions (Fig. 1). This suggests that this method of predicting novel trophic interactions can be used with food webs that have been relatively poorly described. Prediction of novel trophic interactions was far more robust to poor sampling than the estimation of various food-web properties, such as food chain length and connectance, which were highly sensitive to unsampled trophic interactions (Goldwasser and Roughgarden 1997, Martinez et al. 1999). Missing information about trophic interactions affected predictions from the trophic model to a lesser degree than erroneous (false positive) records of interactions. This suggests that in the compilation of food webs, it may be advisable to exclude records of dubious quality rather than include them.

Analogous recent work has explored how sampling affects the analysis of environmental niche models (ENMs; Peterson et al. 2011), and it is worthwhile to compare that work with our results from food-web models. In general, ENMs appear robust to poorly sampled occurrence records of organisms (Stockwell and

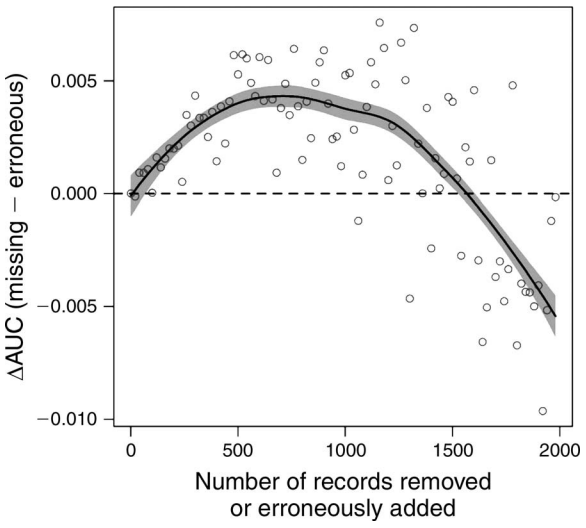

FIG. 2. The relative cost of missing vs. erroneous records. A comparison of the effect of missing vs. erroneous records on the predictiveness (AUC) of a trophic model onto nonnative Lepidoptera–plant interactions; AUC of model with erroneous records is subtracted from AUC of model with missing records. Values of ΔAUC above 0 indicate that erroneous records have a greater cost to model predictiveness than missing records. Open circles represent the difference in means of five simulations with missing data and five simulations with erroneous data. The line is a local-fitted polynomial spline, and shading represents the standard error.

Peterson 2002, Wisz et al. 2008), much as our model suggests robustness to undersampling of interaction records. For example, most algorithms used in ENMs achieved 90% accuracy of within-sample predictions with only 10 occurrence records of Mexican birds (Stockwell and Peterson 2002). In this case, relatively few records were necessary to delimit the set of environmental parameters that define a bird's environmental niche. In our case, interactions with only a subset of native hosts were adequate to define the phylogenetic groupings of hosts that were consumed by an herbivore. One key difference between ENMs and host-use models is that in the latter case, information about native hosts is used twice (in model parameterization and extrapolation), while in ENMs, native occurrence records are used only in model parameterization. We found that the model parameterization phase of host-use models was less sensitive to missing native host records than the extrapolation phase (Fig. 4). In both ENMs and host-use models, there are likely biases in which records go unsampled. In the case of ENMs, records are likely skewed to reflect where biologists tend to collect organisms (Kadmon et al. 2004), and in the case of host-use models, interactions that are either common or apparent are more likely to be sampled (Southwood and Henderson 1966). However, we found that disproportionately removing interactions with lower encounter frequencies had little effect on the predictive ability of our host-use model (Fig. 1A), indicating that this bias may not affect the predictiveness/sensitivity of novel host-use models.

Host-use models, as envisioned here, will be most useful for predicting novel interactions at a regional scale, where large food webs can be compiled. Food webs at this scale have been termed metawebs (Dunne 2006), because they consider interactions over broad spatial and temporal scales. This contrasts food webs often considered in analyses of local communities, where interactions are likely occurring at the same time and within a small area (Elias et al. 2013). At smaller scales, local processes such as competitive exclusion, differential predation, and apparent competition may affect the host affiliations of herbivores in addition to phylogenetic constraints of host use. Host-use models may, however, be useful in defining which interactions are possible within a local food web, though other factors may also inhibit an interaction from being realized. From a practical standpoint, regional metawebs are ideal for the prediction of novel interactions, because the regional scale (i.e., province, state, nation) is the scale at which introduced species are typically managed or quarantined (NAS 2002), so this is the scale at which host-use predictions might be most useful.

Another key topic with host-use models is the taxonomic and ecological scope at which they are useful. Currently, we have shown that host-use models based on phylogenetic proximity and number of hosts are accurate in out-of-sample predictions of hosts of herbivorous Lepidoptera. The high predictive power of phylogenetic proximity of hosts is perhaps unsurprising when considering host affiliations of fairly specialized

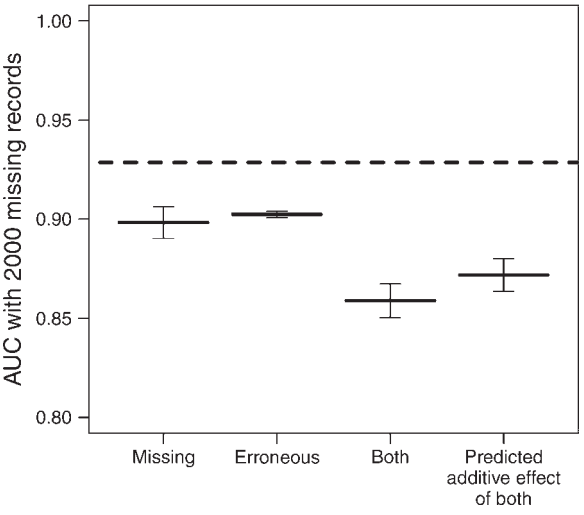

FIG. 3. Additivity of novel host predictions with missing and erroneous records. The effect of 2000 missing, 2000 erroneous, and both (2000 missing + 2000 erroneous) records on the prediction (AUC) of novel Lepidoptera–plant interactions. The predicted additive effect of both missing and erroneous data was calculated. The dashed line indicates the predictiveness (AUC) of the model with the full data set. Bars are means from 20 bootstrap simulations ± standard deviation.

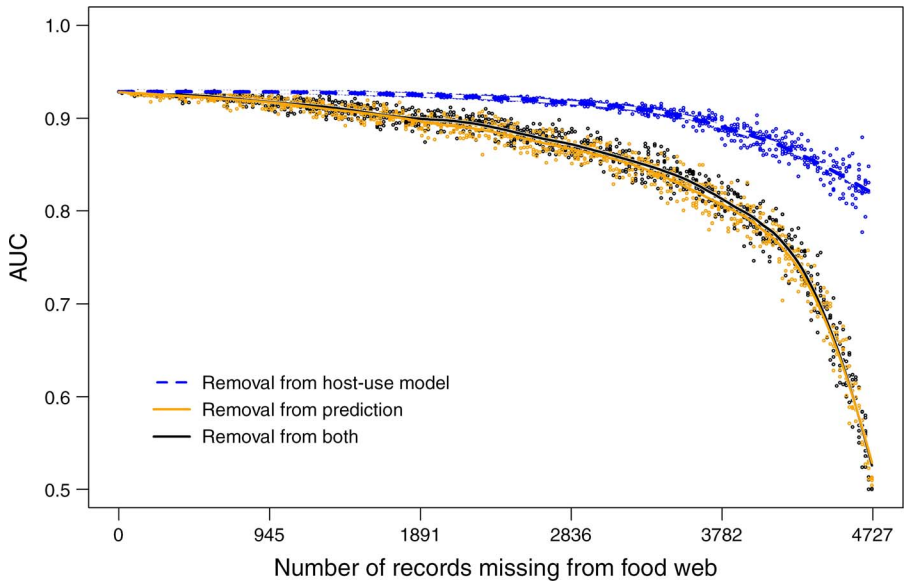

FIG. 4. Records missing (i.e., removed) at different steps of the predictive model. Out-of-sample predictions (AUC) of trophic models onto nonnative Lepidoptera interactions with increasing numbers of records removed from the native food web; numbers are shown for 0%, 20%, 40%, 60%, 80%, and 100% of records removed. Records were removed either from all parts of the predictive model (black), removed in the parameterization phase of the host-use model (blue), or removed from the prediction phase (orange). Data points represent a single simulation. Lines are local-fitted polynomial splines.

herbivores such as many Lepidoptera, as it has long been known that these herbivores consume a phylogenetically defined set of hosts (Ehrlich and Raven 1964, Connor et al. 1980, Futuyma 1983, Ødegaard et al. 2005, Weiblen et al. 2006). Many other interactions, such as host–parasite interactions (Ives and Godfray 2006), fungal pathogen interactions with plants (Gilbert et al. 2012), and predator–prey interactions (Naisbit et al. 2012) are constrained to varying degrees by phylogenetic proximity of hosts. The phylogenetic signal in these interactions may translate to higher trophic levels in some cases (Leppänen et al. 2013), but not others (Elias et al. 2013). For interactions defined along other axes, host information beyond phylogenetic proximity will likely be necessary to make accurate predictions. For example, body size relationships predicted trophic interactions in a food web of Mediterranean fish (Gravel et al. 2013). Even for plant–herbivore interactions, not only phylogenetic proximity, but also leaf trait similarity explained variation in herbivore damage to nonnative oak trees (Pearse and Hipp 2009). In the same system, the nonnative hosts of a polyphagous herbivore were defined by leaf defensive traits irrespective of their similarity to a local native (Pearse 2011).

Currently, our host-use model uses only basic modeling approaches (GLMs) and very little information about the organisms involved. This simplicity has some advantages. For example, while phylogenetic relationships can be estimated from the literature for most plants, their relevant defenses against herbivores cannot. It is likely, however, that including more information about interacting organisms and using more sophisticated modeling techniques that can fit more complex interactions will improve the predictive ability of host-use models even further. Drawing another analogy to environmental niche models, the inclusion of multiple environmental parameters and the use of sophisticated algorithms such as Maxent and GARP consistently improve ENMs over simpler models (Wisz et al. 2008, Elith and Leathwick 2009).

How complete are native food webs?—We found that host-use models retained high predictive ability of novel interactions until roughly two-thirds of all trophic interactions were removed from the native food web. If the same pattern is true for other food webs, it would suggest that a one-third sampling completeness is necessary for host-use models to be useful. This begs the question: how well-sampled are trophic interactions in various food webs? Sampling likely varies widely among food webs, though this can be difficult to determine quantitatively because of the heterogeneous way in which most large food webs are necessarily compiled. For example, in one plant–pollinator food web, intense sampling using traditional direct observation of pollinator visits to flowers missed 26% of all pollinator visitation links, which were later confirmed using pollen fingerprinting methods (Olesen et al. 2011). Using insect–plant food webs as an example, the food webs with the highest sampling intensity tend to be confined to a particular region or location. For example, the host plants of most British butterflies and moths are well-described (e.g., Dennis et al. 2004), and the hosts of tropical herbivorous insects are well-studied for a few geographically limited locations (e.g., Weiblen et al.

2006), but are largely unknown for large tropical areas. In contrast, large-scale monographs and databases of insect–plant interactions (the USDA APHIS-PPQ [Plant Protection and Quarantine] Global Pest and Disease Database, Tietz 1972, Robinson et al. 2010) probably represent a very small fraction of the total host range of those insects. These are the resources, however, that will be most applicable to predicting important novel arthropod–plant interactions, including herbivory to nonnative plants (Gilbert et al. 2012), and nontarget effects of biological control agents (Louda et al. 2003, Desurmont and Pearse 2014). Using the latter as an example, intensive, small-scale laboratory studies are currently being conducted to anticipate nontarget effects of biological control agents, but these tests are costly and occasionally fail to anticipate novel hosts, with disastrous consequences (Louda et al. 2003). Food-web modeling approaches could complement feeding studies to provide a more complete assessment of potential nontarget hosts. Fortunately, while increased sampling effort of native food webs will likely make those food webs slightly more useful for inferring potential novel interactions with introduced species, we show that poor sampling does not necessarily impede those predictions.

Acknowledgments

We would like to thank Claudio Gratton, David Zaya, Samantha Primer, Brenda Molano-Flores, and two anonymous reviewers for useful discussions and comments that improved our manuscript.

Literature Cited

Altermatt, F. 2010. Tell me what you eat and I'll tell you when you fly: diet can predict phenological changes in response to climate change. Ecology Letters 13:1475–1484.

Altermatt, F., D. Fritsch, W. Huber, and S. Whitebread. 2006. Die Gross-Schmetterlingsfauna der Region Basel. Entomologische Gesellschaft Basel, Basel, Switzerland.

Altermatt, F., and I. S. Pearse. 2011. Similarity and specialization of the larval versus adult diet of European butterflies and moths. American Naturalist 178:372–382.

Bremer, B., et al. 2009. An update of the Angiosperm Phylogeny Group classification for the orders and families of flowering plants: APG III. Botanical Journal of the Linnean Society 161:105–121.

Briese, D. T. 2003. The centrifugal phylogenetic method used to select plants for host-specificity testing of weed biological control agents: can and should it be modernised? Pages 23–33 *in* J. H. Spafford and D. T. Briese, editors. CRC for Australian Weed Management No. 7. CRC for Australian Weed Management, Glen Osmond, Australia.

Bundesamt für Naturschutz. 2010. FloraWeb. Bundesamt für Naturschutz, Bonn, Switzerland. http://www.floraweb.de/index.html

Cattin, M. F., L. F. Bersier, C. Banasek-Richter, R. Baltensperger, and J. P. Gabriel. 2004. Phylogenetic constraints and adaptation explain food-web structure. Nature 427:835–839.

Connor, E. F., S. H. Faeth, D. Simberloff, and P. A. Opler. 1980. Taxonomic isolation and the accumulation of herbivorous insects: a comparison of introduced and native trees. Ecological Entomology 5:205–211.

Dennis, R. L. H., J. G. Hodgson, R. Grenyer, T. G. Shreeve, and D. B. Roy. 2004. Host plants and butterfly biology: do host-plant strategies drive butterfly status? Ecological Entomology 29:12–26.

Desurmont, G. A., and I. S. Pearse. 2014. Alien plants versus alien herbivores: does it matter who is non-native in a novel trophic interaction? Current Opinion in Insect Science 2:20–25.

Dunne, J. A. 2006. The network structure of food webs. Pages 27–86 *in* M. Pascual and J. A. Dunne, editors. Ecological networks: linking structure and dynamics. Oxford University Press, Oxford, UK.

Durka, W., and S. G. Michalski. 2012. Daphne: a dated phylogeny of a large European flora for phylogenetically informed ecological analyses. Ecology 93:2297–2297.

Ebert, G., editor. 1991–2005. Die Schmetterlinge Baden-Württembergs. Ulmer, Stuttgart, Germany.

Ehrlich, P. R., and P. H. Raven. 1964. Butterflies and plants: a study in coevolution. Evolution 18:586–608.

Elias, M., C. Fontaine, and F. J. F. van Veen. 2013. Evolutionary history and ecological processes shape a local multilevel antagonistic network. Current Biology 23:1355–1359.

Elith, J., and J. R. Leathwick. 2009. Species distribution models: ecological explanation and prediction across space and time. Annual Review of Ecology, Evolution, and Systematics 40:677–697.

Futuyma, D. J. 1983. Evolutionary interactions among herbivorous insects and plants. Pages 207–231 *in* D. J. Futuyma and S. M., editors. Coevolution. Sinauer, Sunderland, Massachusetts, USA.

Gilbert, G. S., R. Magarey, K. Suiter, and C. O. Webb. 2012. Evolutionary tools for phytosanitary risk analysis: phylogenetic signal as a predictor of host range of plant pests and pathogens. Evolutionary Applications 5:869–878.

Goldwasser, L., and J. Roughgarden. 1997. Sampling effects and the estimation of food-web properties. Ecology 78:41–54.

Gómez, J. M., M. Verdú, and F. Perfectti. 2010. Ecological interactions are evolutionarily conserved across the entire tree of life. Nature 465:918–U916.

Gravel, D., T. Poisot, C. Albouy, L. Velez, and D. Mouillot. 2013. Inferring food web structure from predator–prey body size relationships. Methods in Ecology and Evolution 4:1083–1090.

Harvey, J. A., et al. 2010. Ecological fits, mis-fits and lotteries involving insect herbivores on the invasive plant, *Bunias orientalis*. Biological Invasions 12:3045–3059.

Ives, A. R., and H. C. J. Godfray. 2006. Phylogenetic analysis of trophic associations. American Naturalist 168:E1–E14.

Kadmon, R., O. Farber, and A. Danin. 2004. Effect of roadside bias on the accuracy of predictive maps produced by bioclimatic models. Ecological Applications 14:401–413.

Koch, M., and W. Heinicke. 1991. Wir bestimmen Schmetterlinge. Ausgabe in einem Band. Neumann/Neudamm, Melsungen, Germany.

Krzanowski, W. J., and D. J. Hand. 2009. ROC curves for continuous data. Chapman & Hall, Boca Raton, Florida, USA.

Leppänen, S. A., E. Altenhofer, A. D. Liston, and T. Nyman. 2013. Ecological versus phylogenetic determinants of trophic associations in a plant–leafminer–parasitoid food web. Evolution 67:1493–1502.

Levine, J. M., P. B. Adler, and S. G. Yelenik. 2004. A meta-analysis of biotic resistance to exotic plant invasions. Ecology Letters 7:975–989.

Louda, S. M., R. W. Pemberton, M. T. Johnson, and P. A. Follett. 2003. Nontarget effects: the Achilles' Heel of biological control? Retrospective analyses to reduce risk associated with biocontrol introductions. Annual Review of Entomology 48:365–396.

Mack, R. N., D. Simberloff, W. M. Lonsdale, H. Evans, M. Clout, and F. A. Bazzaz. 2000. Biotic invasions: causes, epidemiology, global consequences, and control. Ecological Applications 10:689–710.

Maron, J. L., and M. Vilá. 2001. When do herbivores affect plant invasion? Evidence for the natural enemies and biotic resistance hypotheses. Oikos 95:361–373.
Martinez, N. D., B. A. Hawkins, H. A. Dawah, and B. P. Feifarek. 1999. Effects of sampling effort on characterization of food-web structure. Ecology 80:1044–1055.
Naisbit, R. E., R. P. Rohr, A. G. Rossberg, P. Kehrli, and L. F. Bersier. 2012. Phylogeny versus body size as determinants of food web structure. Proceedings of the Royal Society B 279:3291–3297.
NAS. 2002. Predicting invasions of nonindigenous plants and plant pests. National Academy Press, Washington, D.C., USA.
Ødegaard, F., O. H. Diserud, and K. Østbye. 2005. The importance of plant relatedness for host utilization among phytophagous insects. Ecology Letters 8:612–617.
Olesen, J. M., J. Bascompte, Y. L. Dupont, H. Elberling, C. Rasmussen, and P. Jordano. 2011. Missing and forbidden links in mutualistic networks. Proceedings of the Royal Society B 278:725–732.
Pearse, I. S. 2011. Leaf defensive traits in oaks and their role in both preference and performance of a polyphagous herbivore, *Orgyia vetusta*. Ecological Entomology 36:635–642.
Pearse, I. S., and F. Altermatt. 2013*a*. Extinction cascades partially estimate herbivore losses in a complete Lepidoptera–plant food web. Ecology 94:1785–1794.
Pearse, I. S., and F. Altermatt. 2013*b*. Predicting novel trophic interactions in a non-native world. Ecology Letters 16:1088–1904.
Pearse, I. S., D. Harris, R. Karban, and A. Sih. 2013. Predicting novel herbivore–plant interactions. Oikos 122:1554–1564.
Pearse, I. S., and A. L. Hipp. 2009. Phylogenetic and trait similarity to a native species predict herbivory on non-native oaks. Proceedings of the National Academy of Sciences USA 106:18097–18102.
Peterson, A. T., J. Soberón, R. G. Pearson, R. P. Anderson, E. Martínez-Meyer, M. Nakamura, and M. B. Araújo. 2011. Ecological niches and geographic distributions. Princeton Press, Princeton, New Jersey, USA.
R Core Team. 2014. R: a language and environment for statistical computing v. 3.1.0. R Foundation for Statistical Computing, Vienna, Austria. www.r-project.org
Rezende, E. L., J. E. Lavabre, P. R. Guimarães, P. Jordano, and J. Bascompte. 2007. Non-random coextinctions in phylogenetically structured mutualistic networks. Nature 448:925–U926.
Robinson, G. S., P. R. Ackery, I. J. Kitching, G. W. Beccaloni, and L. M. Hernández. 2010. HOSTS— A database of the world's lepidopteran hostplants. Natural History Museum, London, UK. http://www.nhm.ac.uk/hosts.
Sih, A., D. I. Bolnick, B. Luttberg, J. L. Orrock, S. D. Peacor, L. M. Pintor, E. Preisser, J. S. Rehage, and J. R. Vonesh. 2010. Predator–prey naïveté, antipredator behavior, and the ecology of predator invasions. Oikos 119:610–621.
Sing, T., O. Sander, N. Beerenwinkel, and T. Lengauer. 2005. ROCR: visualizing classifier performance in R. Bioinformatics 21:3940–3941.
Southwood, T. R. E., and P. A. Henderson. 1966. Ecological methods with particular reference to the study of insect populations. Chapman & Hall, New York, New York, USA.
Stockwell, D. R. B., and A. T. Peterson. 2002. Effects of sample size on accuracy of species distribution models. Ecological Modelling 148:1–13.
Tietz, H. M. 1972. An index to the described life histories, early stages and hosts of the Macrolepidoptera of the continental United States and Canada. A.C. Allyn, Sarasota, Florida, USA.
Verhoeven, K. J. F., A. Biere, J. A. Harvey, and W. H. van der Putten. 2009. Plant invaders and their novel natural enemies: who is naive? Ecology Letters 12:107–117.
Weiblen, G. D., C. O. Webb, V. Novotny, Y. Basset, and S. E. Miller. 2006. Phylogenetic dispersion of host use in a tropical insect herbivore community. Ecology 87:S62–S75.
Wisz, M. S., R. J. Hijmans, J. Li, A. T. Peterson, C. H. Graham, A. Guisan, and NCEAS Predicting Species Distributions Working Group. 2008. Effects of sample size on the performance of species distribution models. Diversity and Distributions 14:763–773.

Supplemental Material

Data Availability

Data associated with this paper have been deposited in Dryad: http://dx.doi.org/10.5061/dryad.5c1g6